Deloitte.

Webinar ''ChatGPT in business''

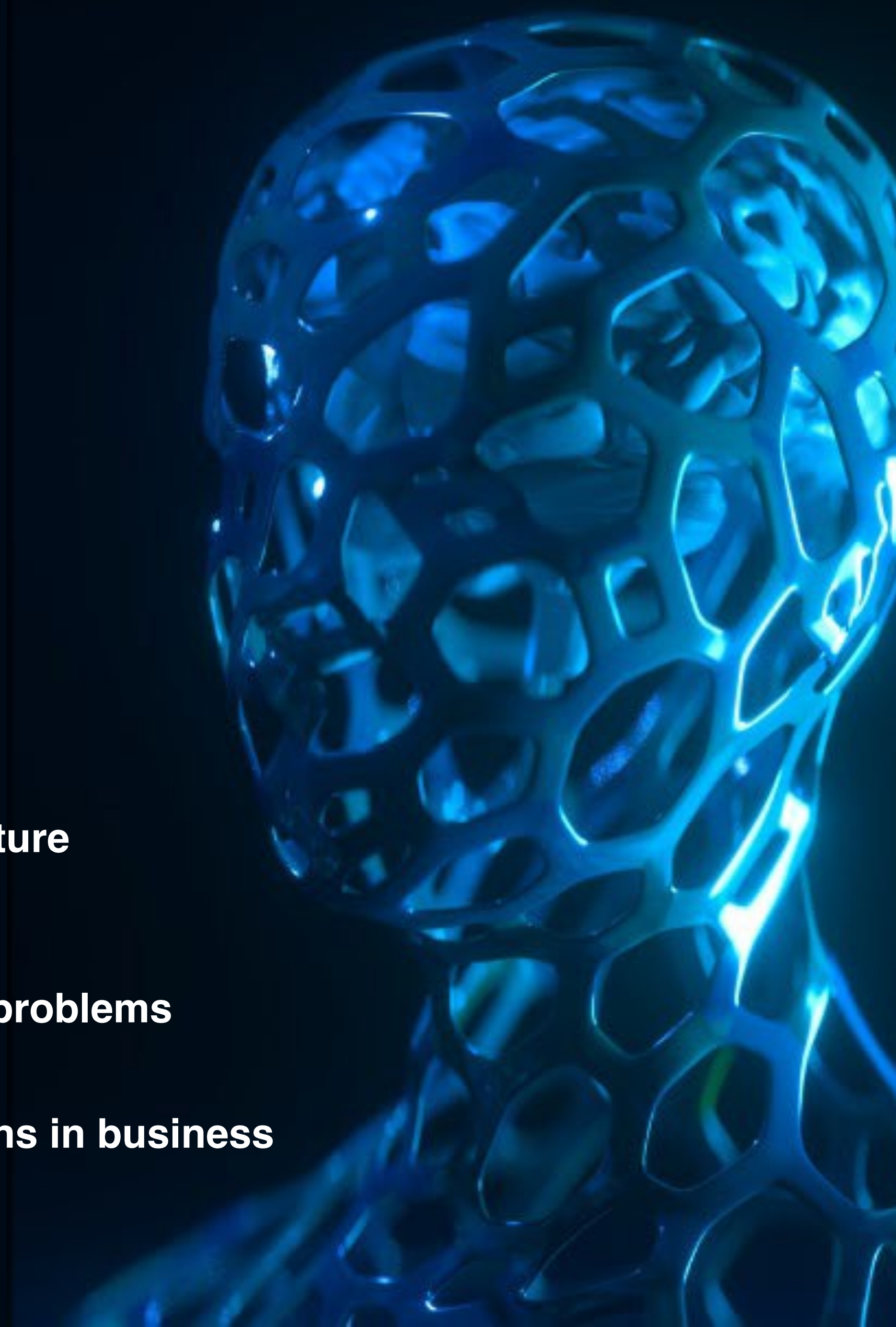
May 15, 2023, 17:00

English

Zoom

Agenda:

- 17:05 – 17:20 I The slow path to intelligence – the current state and future prospects of large language models
  Dr. Ēvalds Urtāns, asya.ai
- 17:20 – 17:35 I Where and when to implement Chat GPT for business problems
  Igor Rodin, Deloitte CE
- 17:35 – 17:50 I Practical guidance on building GPT-4 based applications in business
  Dr. Romāns Taranovs, Deloitte CE



asya

Deloitte webinar
The slow path to intelligence

Evalds Urtans
asya.ai
CEO

What is AI?

Okay gang, let's see who the advanced AI really is.
IF IF IF IF IF IF IF IF IF IF IF IF IF IF IF IF IF IF IF IF IF IF IF

What is AI?

Okay gang, let's see who the advanced AI really is.
IF
IF IF IF
IF IF IF IF
IF IF IF IF IF
IF IF IF IF
IF IF IF IF
IF IF IF IF

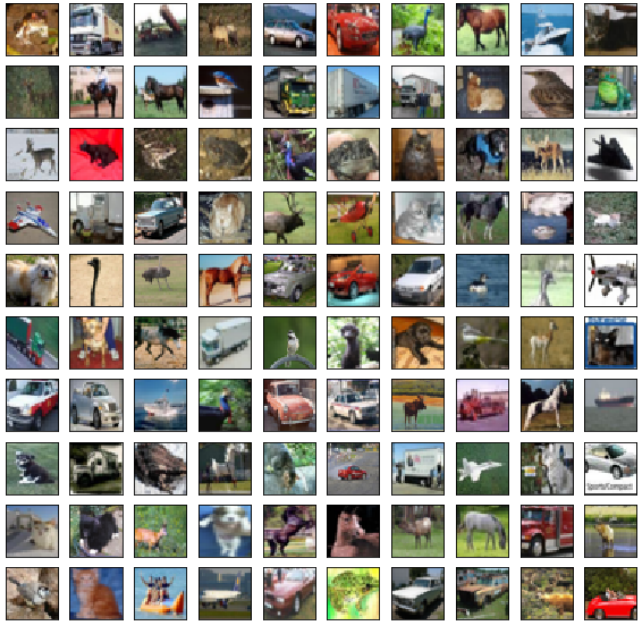
$$
\begin{aligned}
i_t &= \sigma(W_i * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_i) \\
f_t &= \sigma(W_f * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_f) \\
e_{t,z} &= V_e \cdot \tanh(W_e * [\mathcal{X}_{t,z}, \mathcal{H}_{t-1}] + b_e) \\
\alpha_{t,z} &= \frac{\exp\left(e_{t,z}\right)}{\sum_{j=1}^{\tau} \exp\left(e_{t,j}\right)} \\
p_t &= \sum_{j=1}^{\tau} \alpha_{t,j} \tilde{\mathcal{X}}_{t,j} \\
n_t &= \sigma(W_n * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_n) \\
g_t &= \tanh(W_g * [p_t, \mathcal{H}_{t-1}] + b_g) \\
\mathcal{C}_t &= f_t \circ \mathcal{C}_{t-1} + i_t \circ a_t + n_t \circ g_t \\
a_t &= \tanh(W_a * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_a) \\
o_t &= \sigma(W_o * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_o) \\
\mathcal{H}_t &= o_t \circ \tanh(\mathcal{C}_t)
\end{aligned}
$$

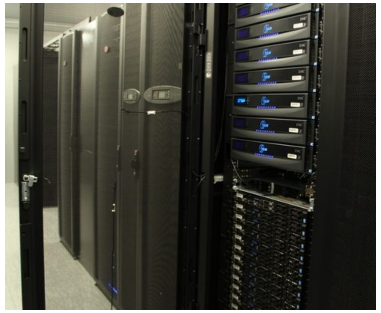
What is AI?

- **Linear algebra**
- **Calculus**
- **Probability theory**
- **Information theory**
- **10% programming**

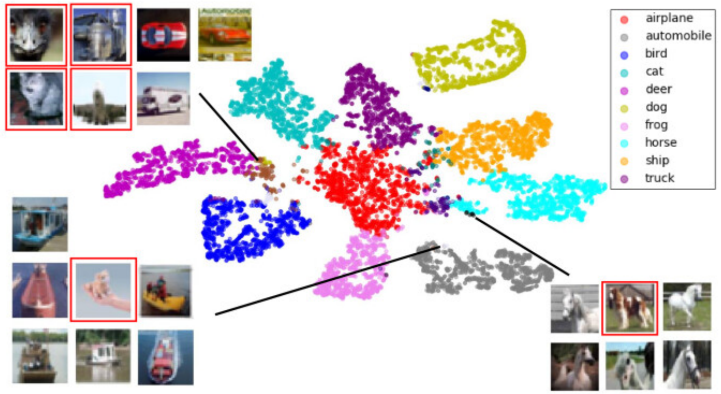
$$
\begin{aligned}
i_t &= \sigma(W_i * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_i) \\
f_t &= \sigma(W_f * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_f) \\
e_{t,z} &= V_e \cdot \tanh(W_e * [\mathcal{X}_{t,z}, \mathcal{H}_{t-1}] + b_e) \\
\alpha_{t,z} &= \frac{\exp\left(e_{t,z}\right)}{\sum_{j=1}^{\tau} \exp\left(e_{t,j}\right)} \\
p_t &= \sum_{j=1}^{\tau} \alpha_{t,j} \tilde{\mathcal{X}}_{t,j} \\
n_t &= \sigma(W_n * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_n) \\
g_t &= \tanh(W_g * [p_t, \mathcal{H}_{t-1}] + b_g) \\
\mathcal{C}_t &= f_t \circ \mathcal{C}_{t-1} + i_t \circ a_t + n_t \circ g_t \\
a_t &= \tanh(W_a * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_a) \\
o_t &= \sigma(W_o * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_o) \\
\mathcal{H}_t &= o_t \circ \tanh(\mathcal{C}_t)
\end{aligned}
$$

What is AI?

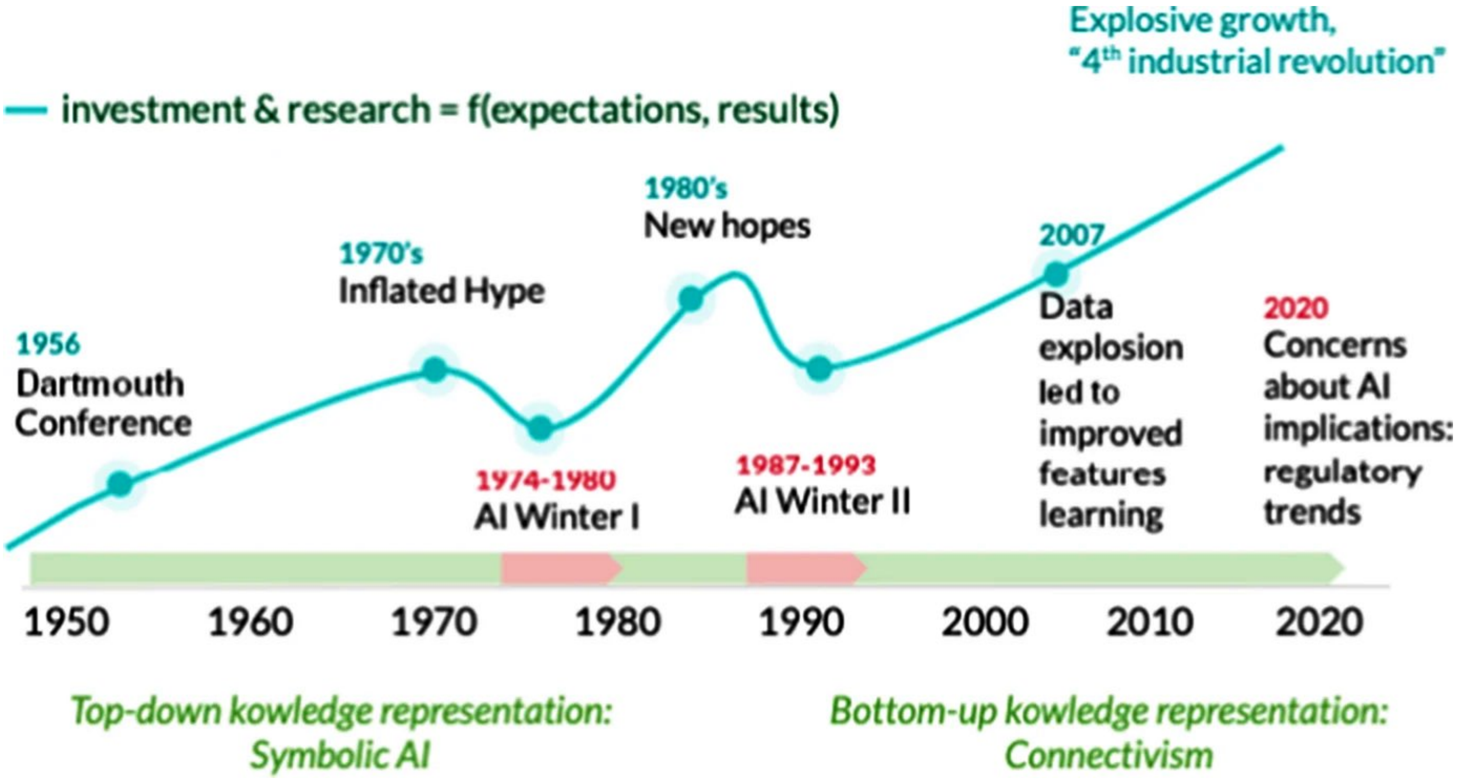




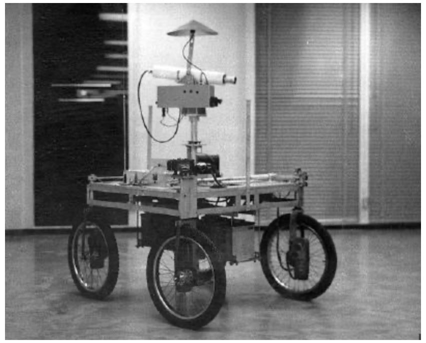
$$
\begin{aligned}
i_t &= \sigma(W_i * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_i) \\
f_t &= \sigma(W_f * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_f) \\
e_{t,z} &= V_e \cdot \tanh(W_e * [\mathcal{X}_{t,z}, \mathcal{H}_{t-1}] + b_e) \\
\alpha_{t,z} &= \frac{\exp\left(e_{t,z}\right)}{\sum_{j=1}^{\tau} \exp\left(e_{t,j}\right)} \\
p_t &= \sum_{j=1}^{\tau} \alpha_{t,j} \tilde{\mathcal{X}}_{t,j} \\
n_t &= \sigma(W_n * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_n) \\
g_t &= \tanh(W_g * [p_t, \mathcal{H}_{t-1}] + b_g) \\
\mathcal{C}_t &= f_t \circ \mathcal{C}_{t-1} + i_t \circ a_t + n_t \circ g_t \\
a_t &= \tanh(W_a * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_a) \\
o_t &= \sigma(W_o * [\mathcal{X}_t, \mathcal{H}_{t-1}] + b_o) \\
\mathcal{H}_t &= o_t \circ \tanh(\mathcal{C}_t)
\end{aligned}
$$



History



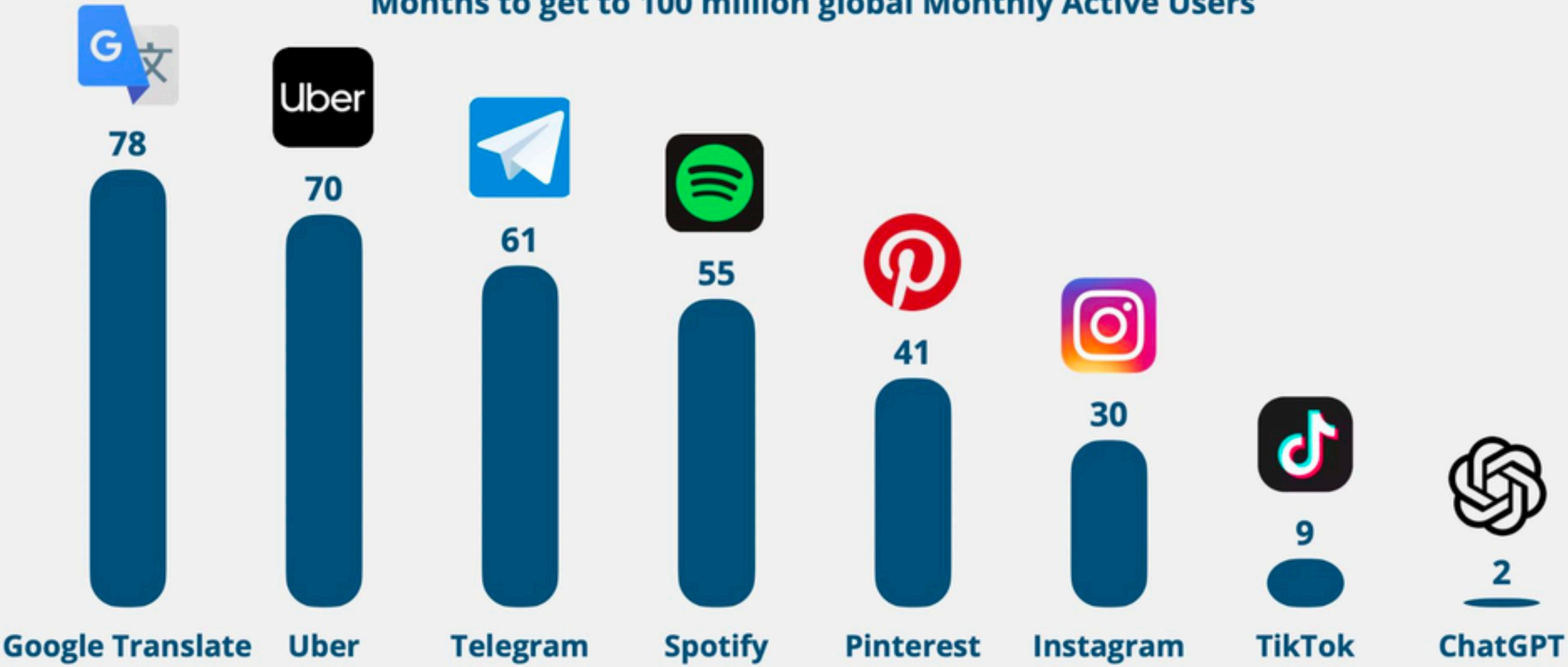
Hans Moravec's Robots, 1975



Tesla FSD, 2023

Time to Reach 100M Users

Months to get to 100 million global Monthly Active Users

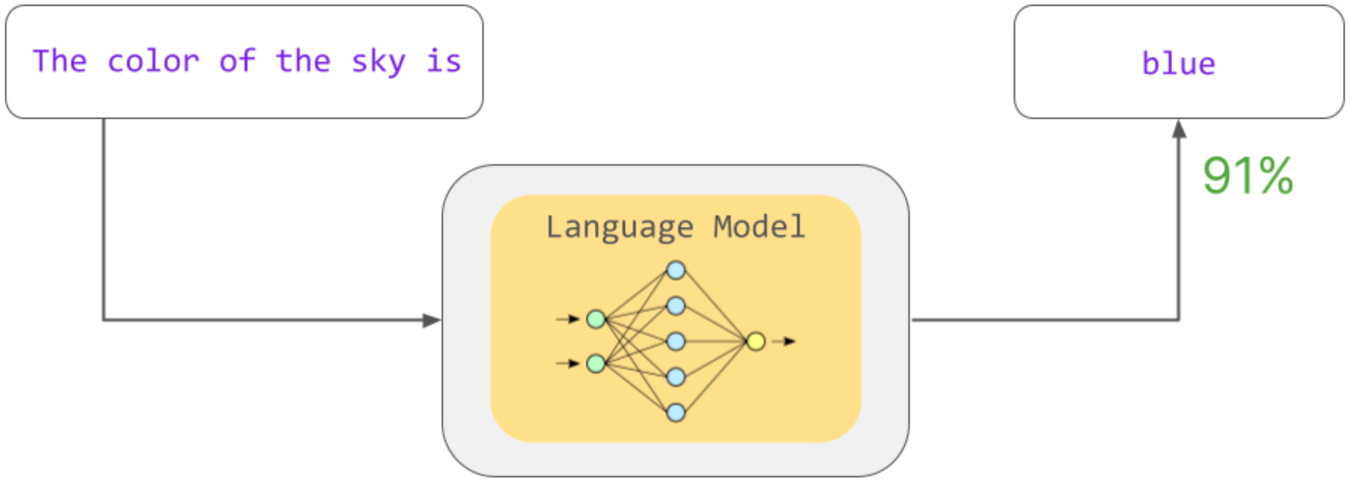


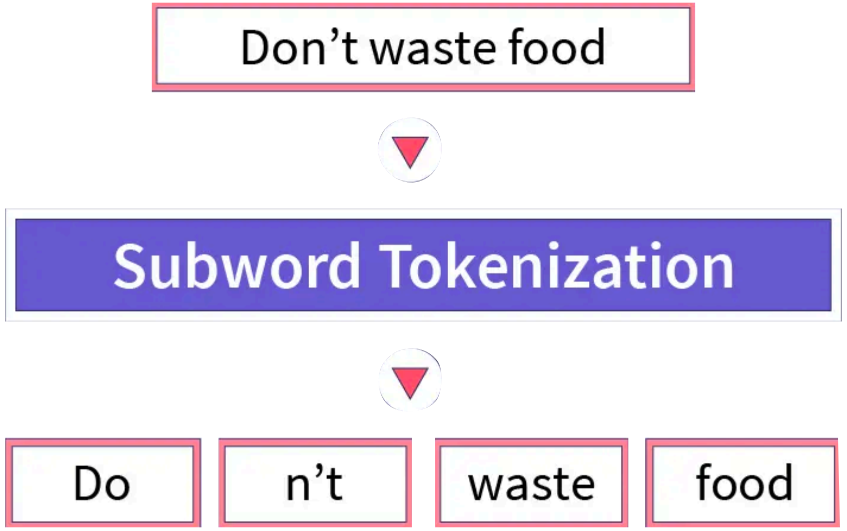
Source: UBS / Yahoo Finance

@EconomyApp

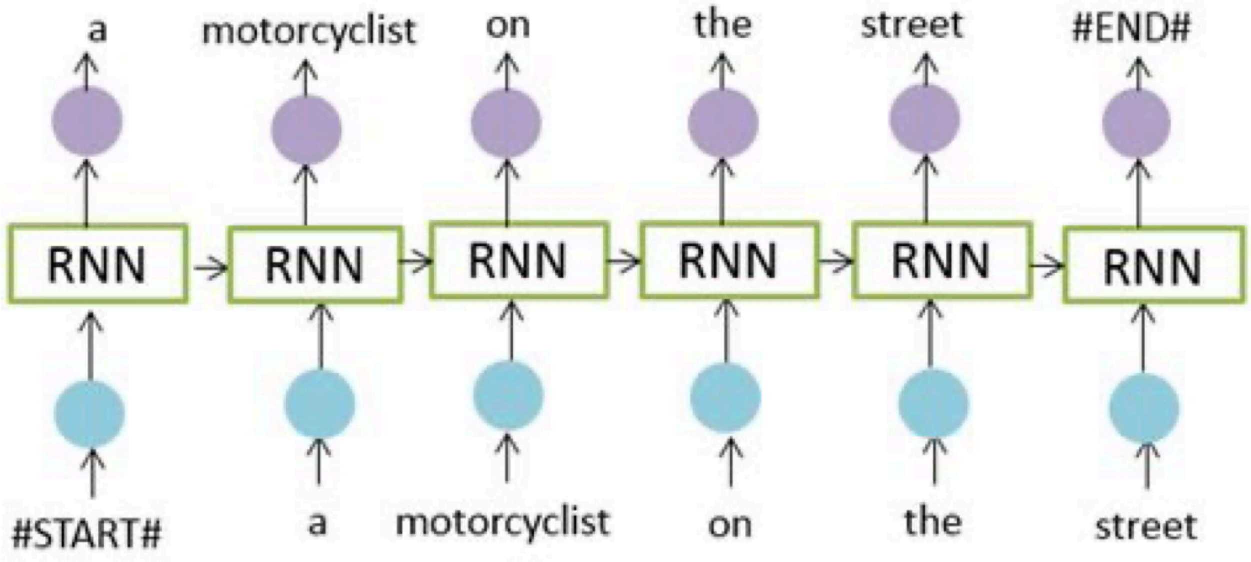
APP ECONOMY **INSIGHTS**

Language modelling, Tokens



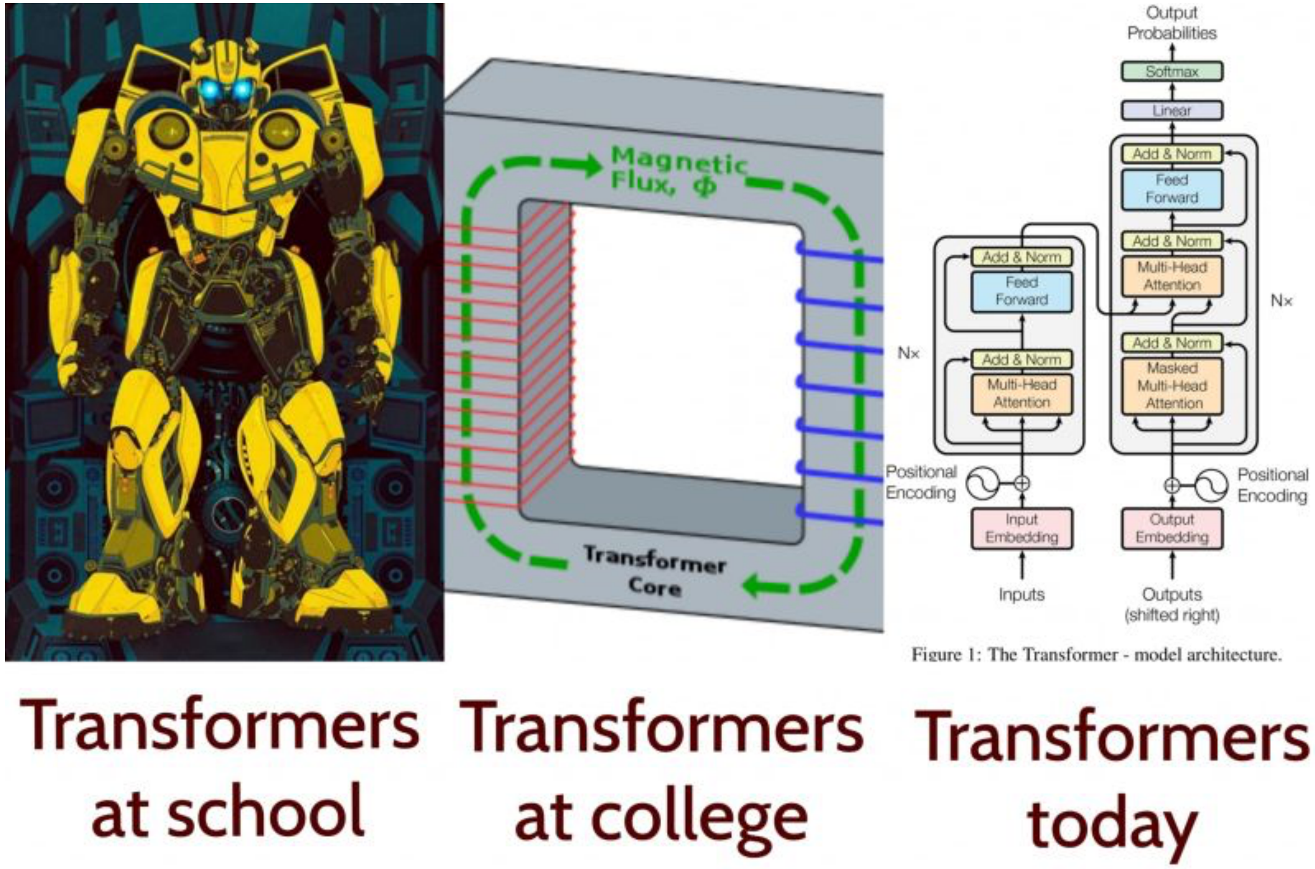


Old way - RNN, LSTM

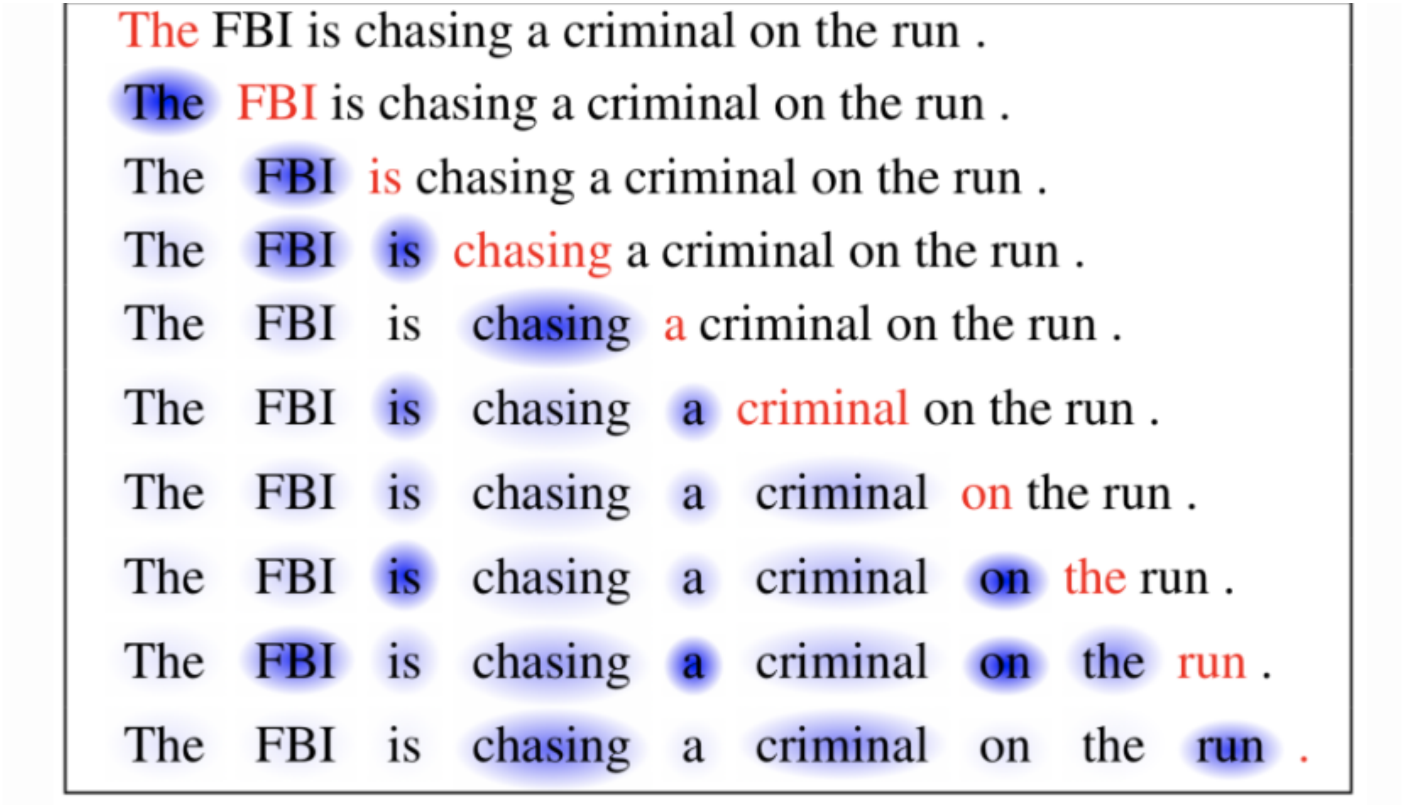


- S. Hochreiter, J. Schmidhuber 1995
- Not-parellizable
- Limited memory capacity
- Small VRAM footprint
- Weaker performance: 74% acc. vs 82% acc (New way)

New way - Transformer



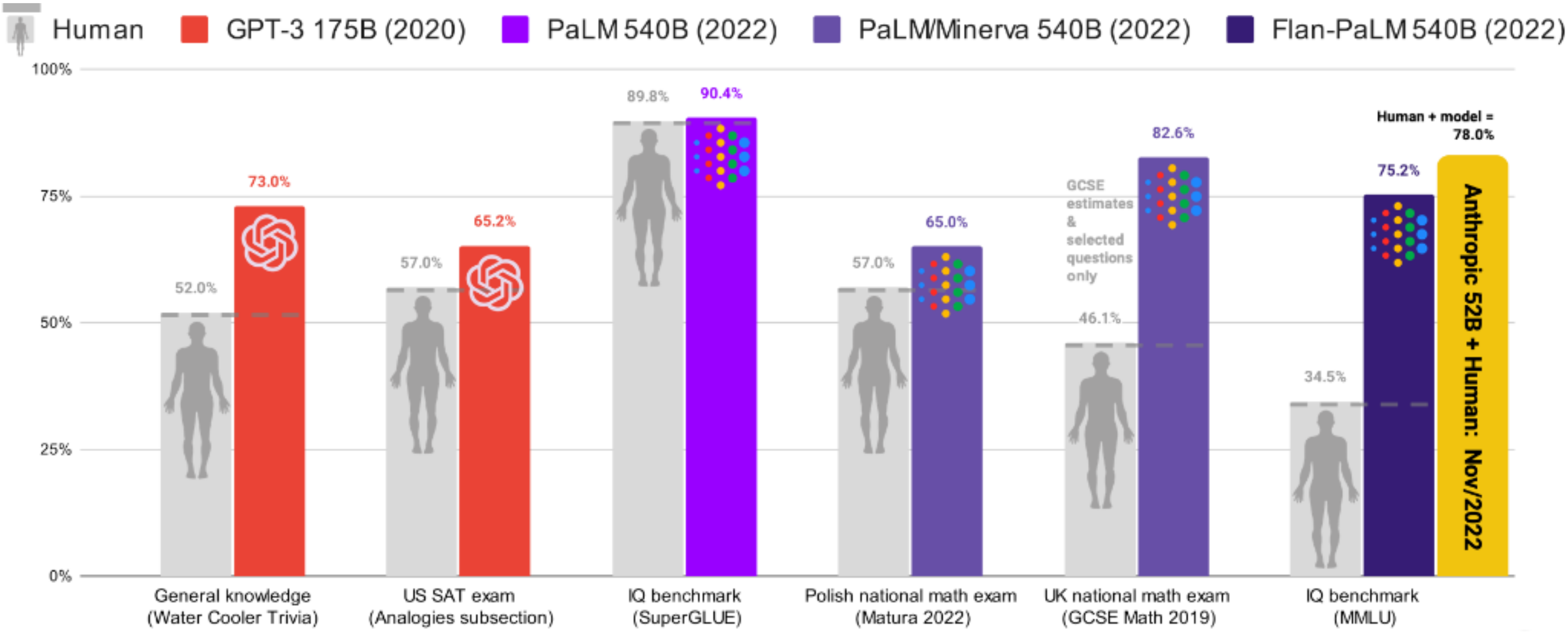
New way - Transformer



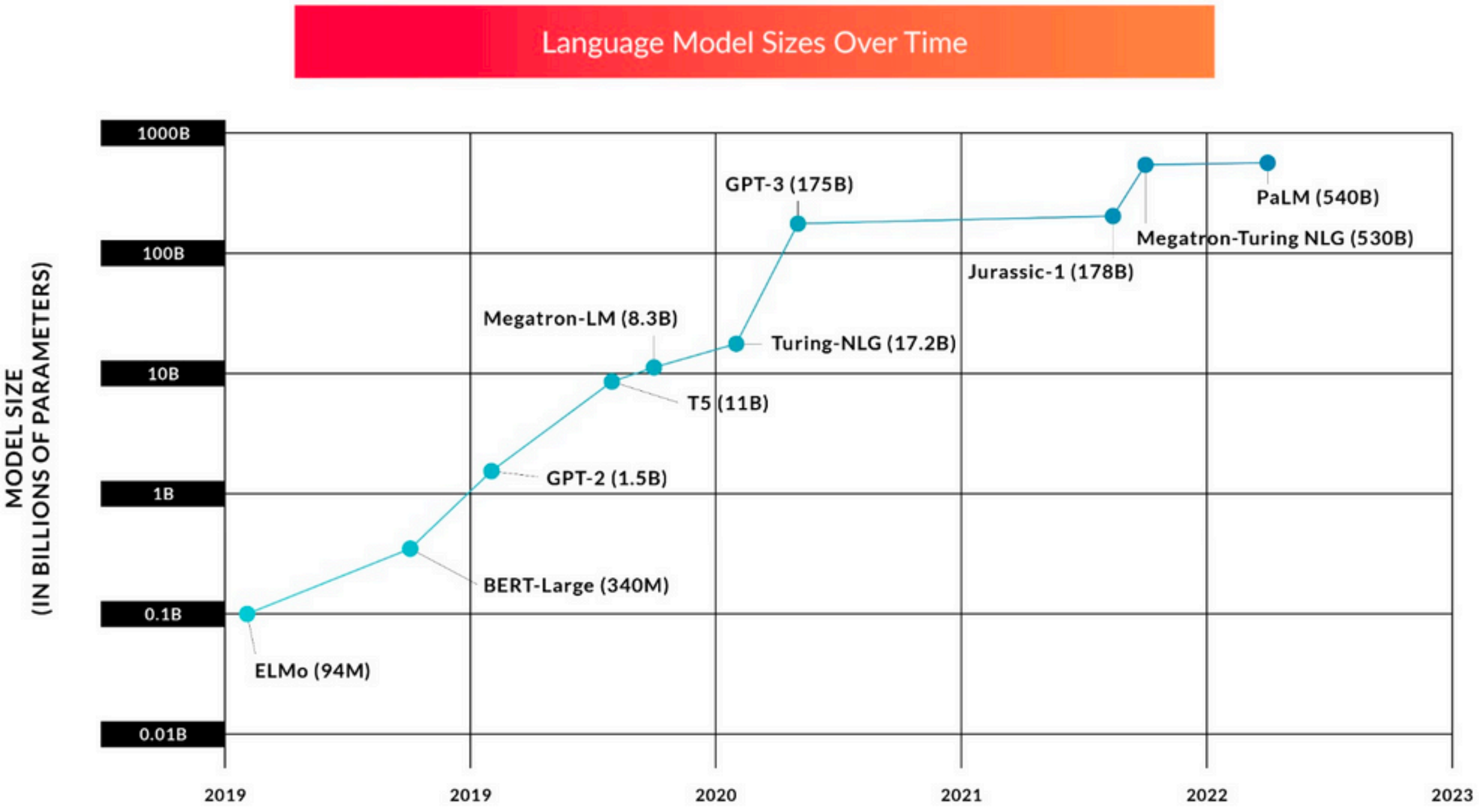
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .
The FBI is chasing a criminal on the run .

- BERT (Google), GPT (OpenAI), 2018
- Transformer architecture
- Parellizable
- No memory*
- Very large VRAM footprint
- Limited context length ~2048 tokens** (or tradeoffs)
- CommonCrawl (10 years), 410b tokens

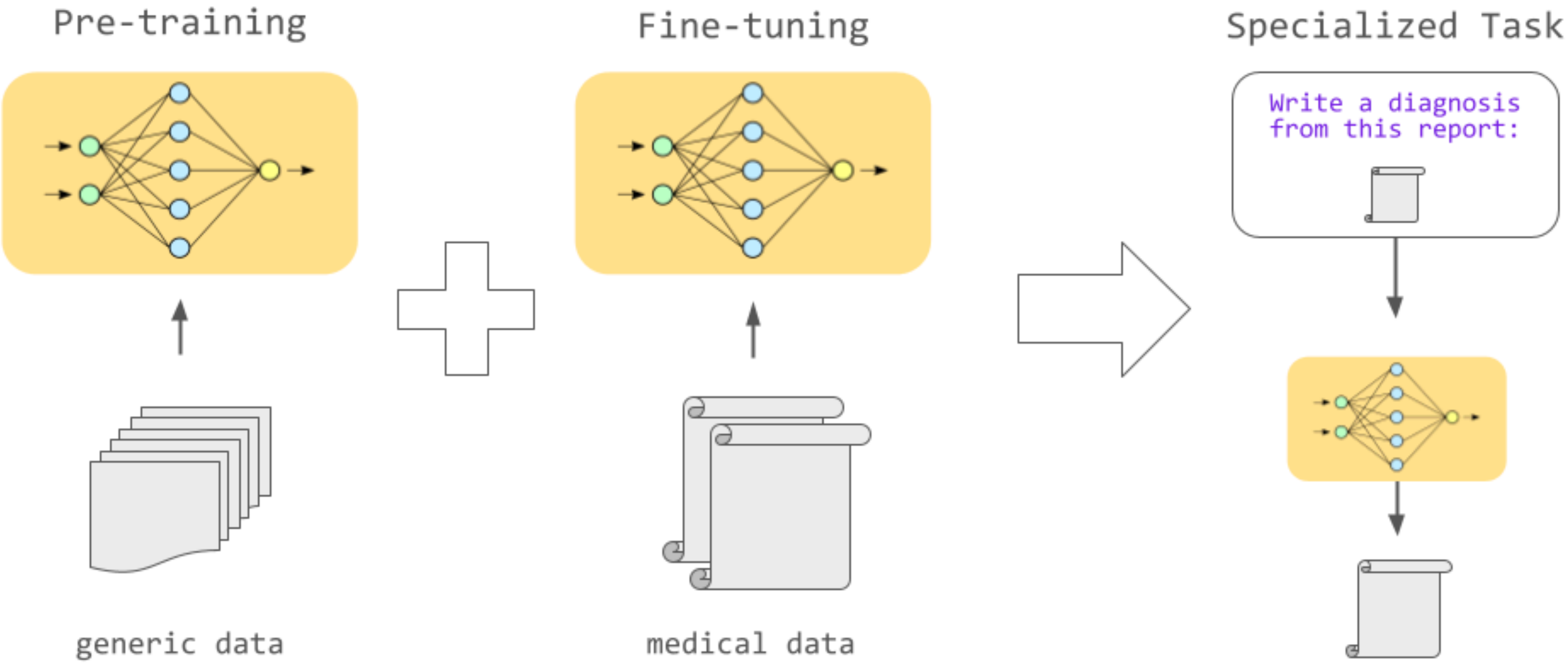
New way - Transformer



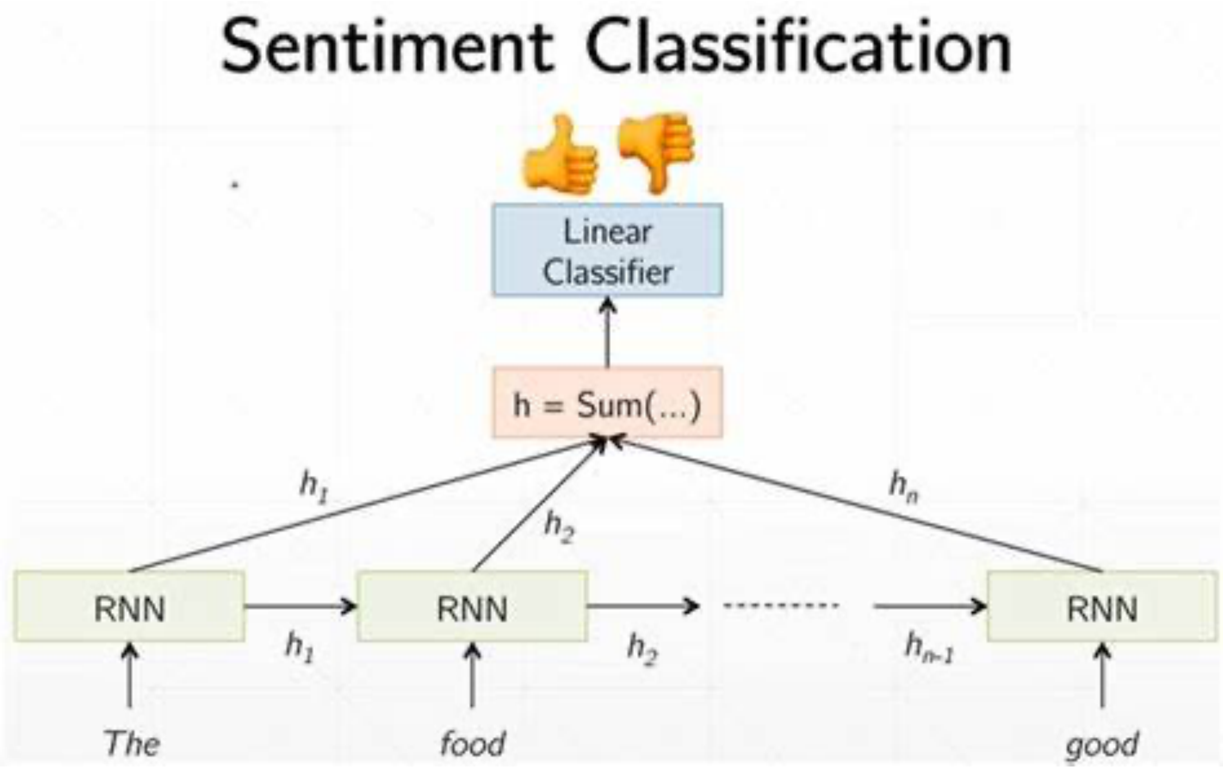
Use-cases - Train from scratch



Use-cases - Fine tunning



Use-cases - Auxilary tasks

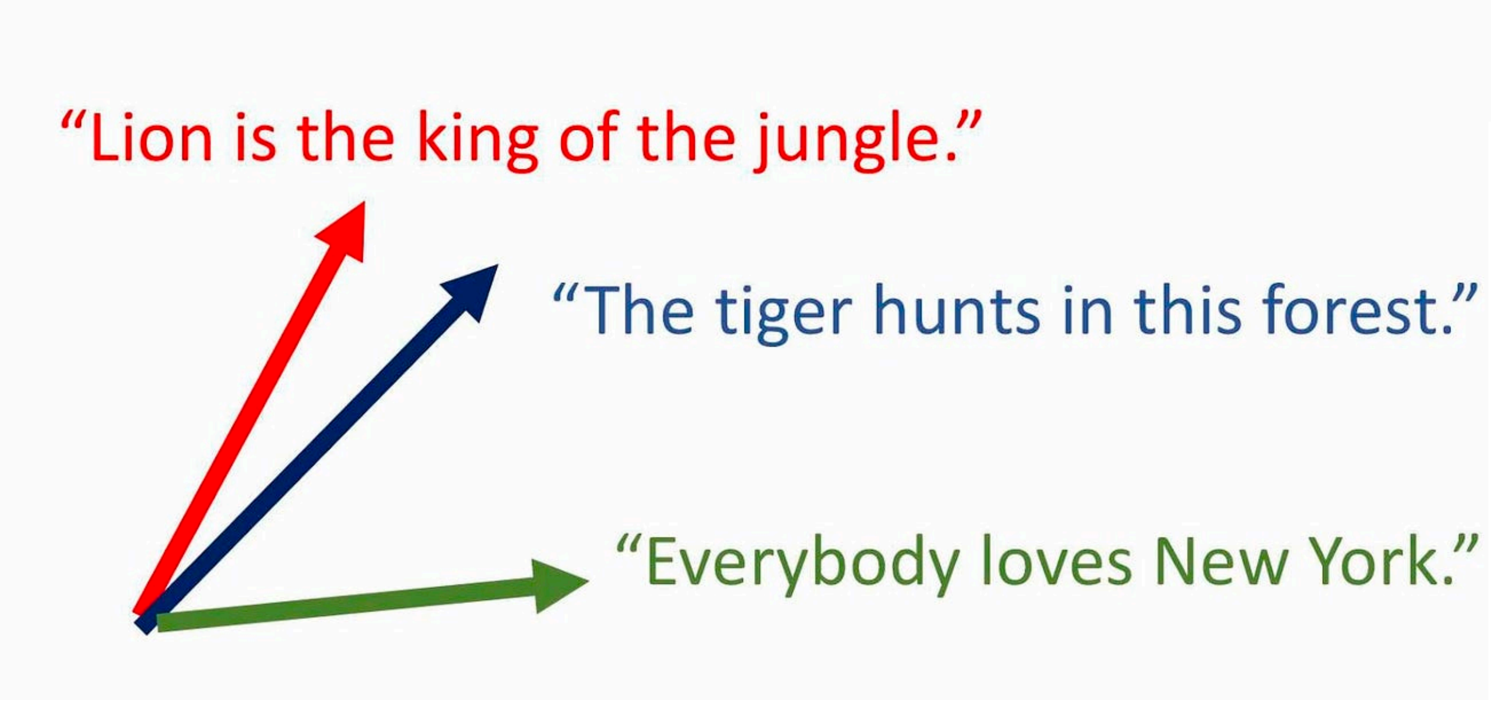


- Classification, Regression, segmentation

- Need labelled data
  at least 10k text input

- Expect higher accuracy 90%+

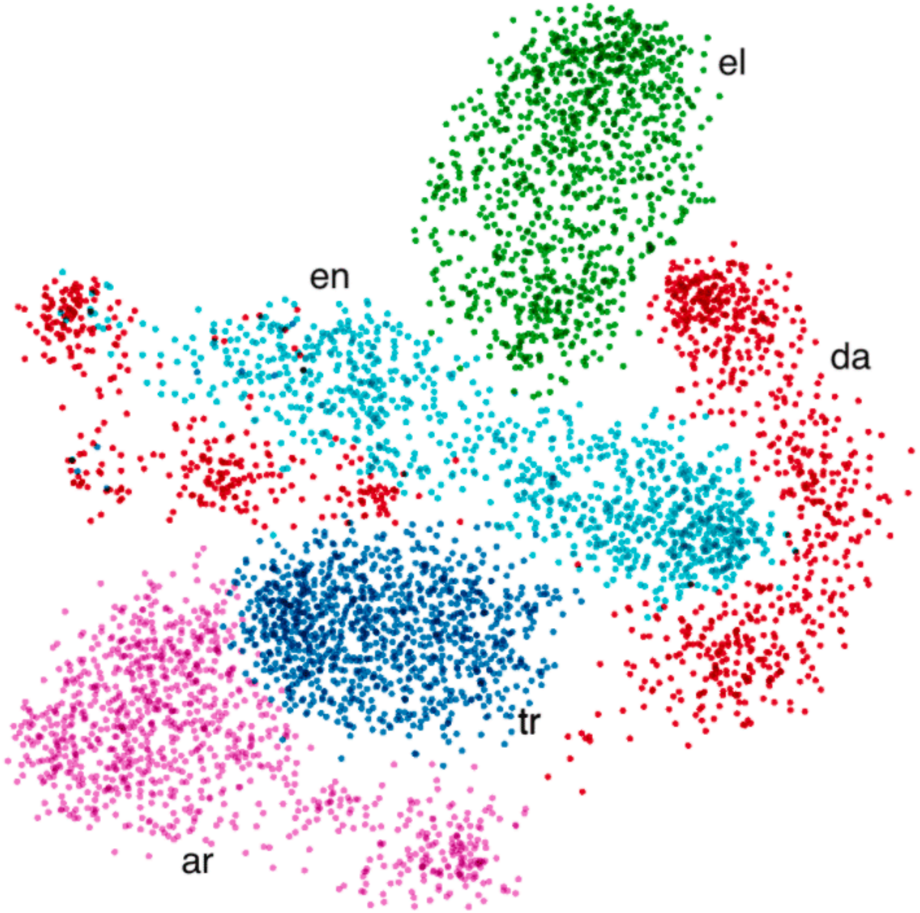
Text embeddings



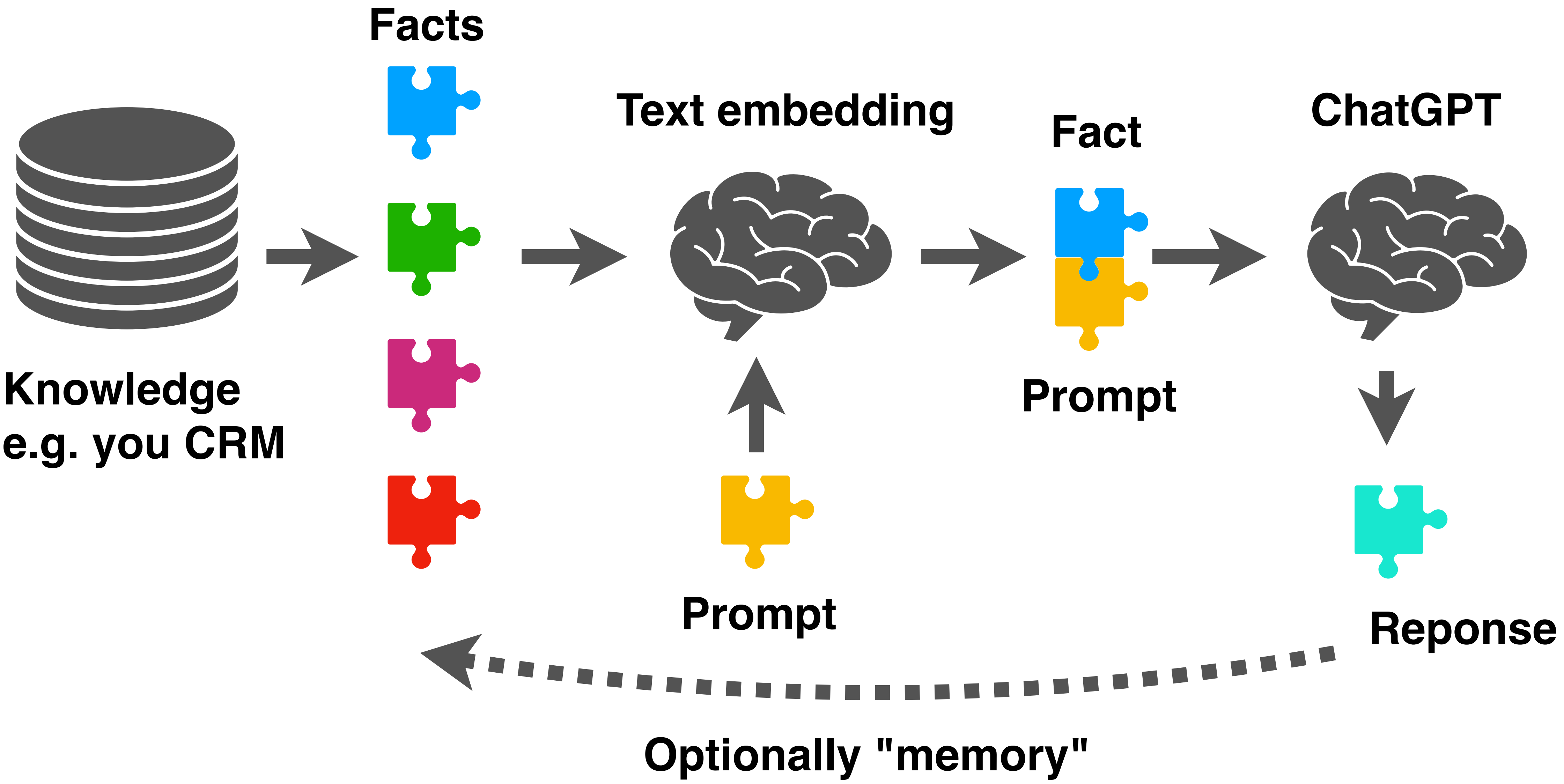
"Lion is the king of the jungle."

"The tiger hunts in this forest."

"Everybody loves New York."



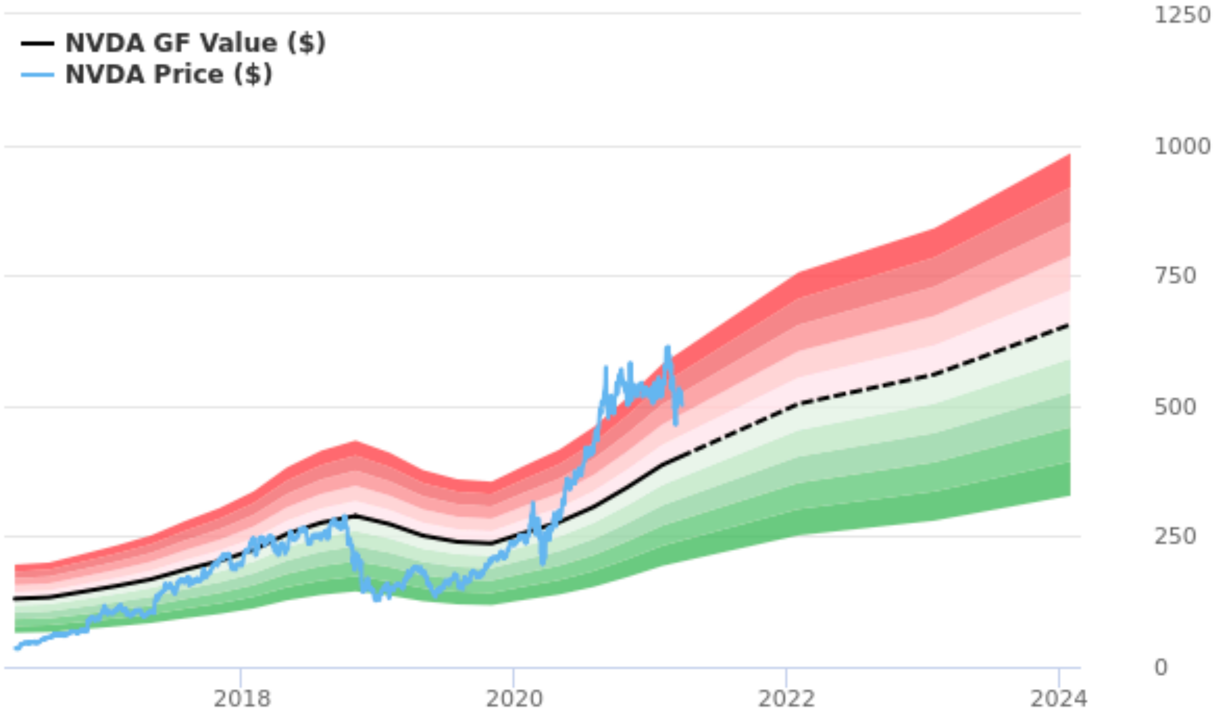
Use-cases - Zero-shot + Knowledge base



Hardware resources

Necessary hardware:

- **AWS EC2, Google Cloud, Oracle cloud - 10k EUR/mon**
- **RTU HPC - <10k EUR/mon**
- **nVidia GPU V100, A100, A10 40GB (min.) - 20k EUR/per unit**
- **Google Collab - 50 EUR/mon**



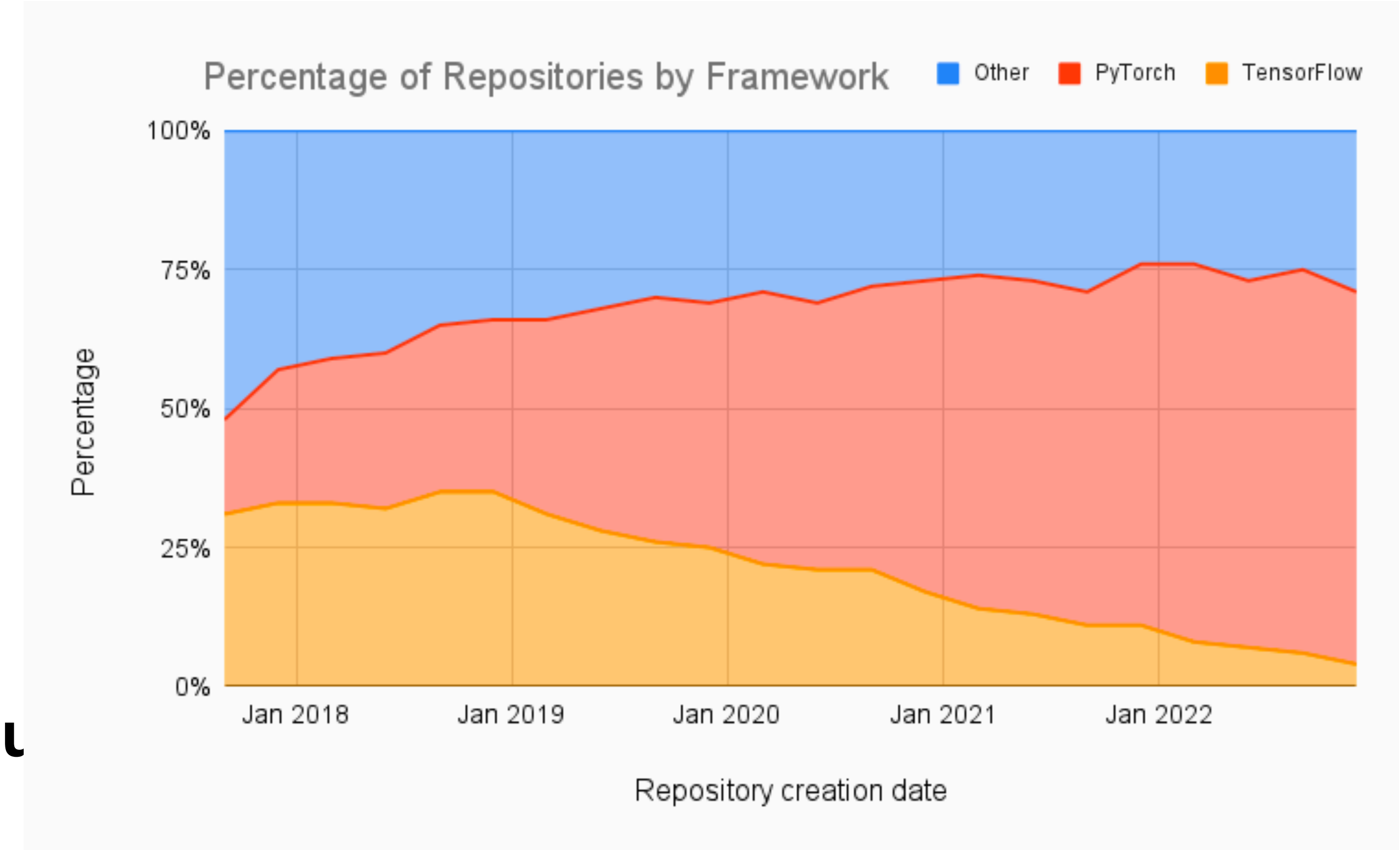
Software resources

Best Open-Source LLMs

- **FLAN-T5 XXL**
- **GPT-JT**
- **Bloom**
- **Open Assistant**
- **LLaMA****

Model sources:

- **huggingface.io**
- **torchvision, torchtext, torchhu**



Model programming:

- **PyTorch**
- **ONNX (cross-platform deployment)**

Tools to explore

Search engines:

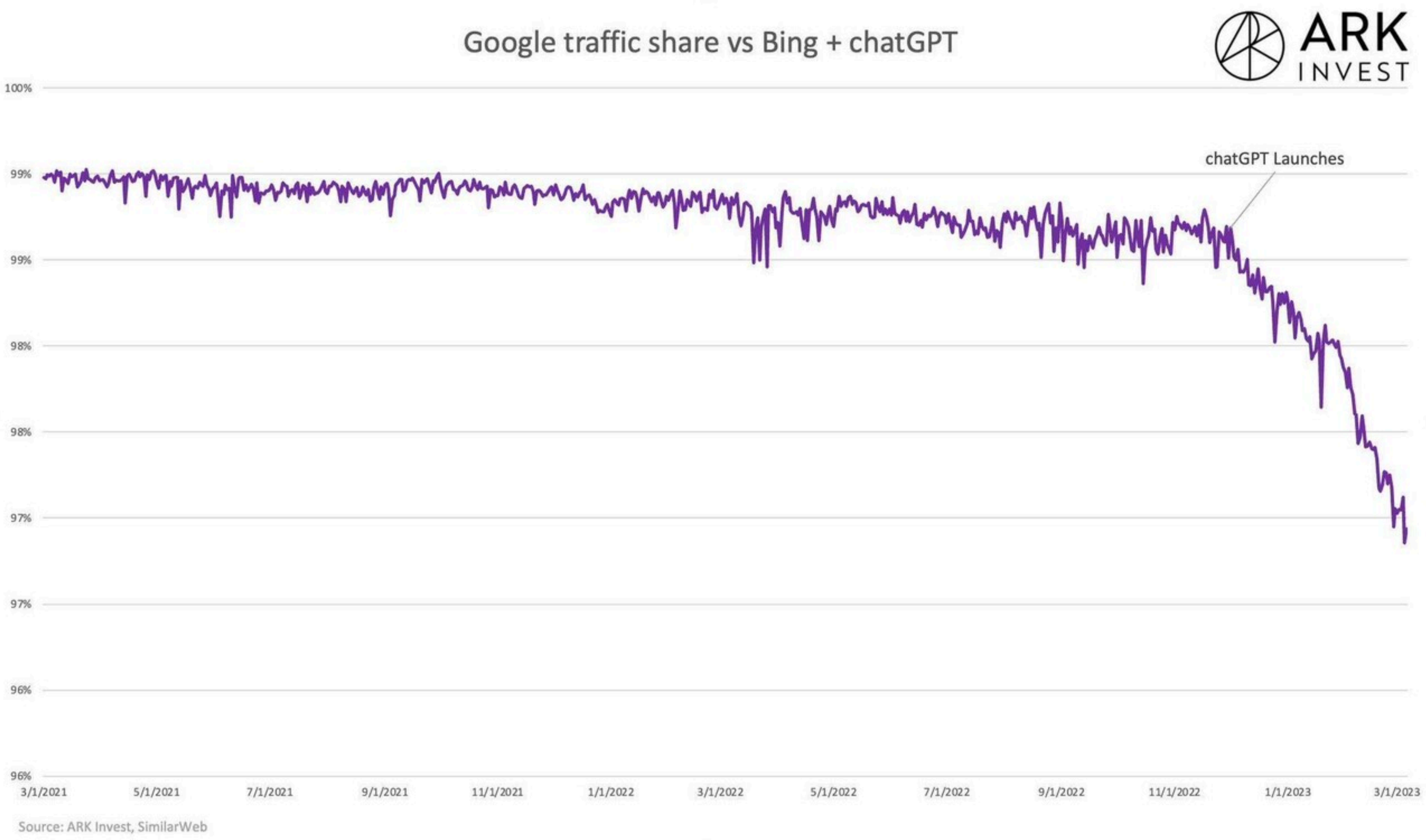
- perplexity.ai
- chat.you.com

Productivity:

- chatpdf.com

Content:

- jasper.ai
- writesonic.ai



Google traffic share vs Bing + chatGPT

chatGPT Launches

Source: ARK Invest, SimilarWeb

Deloitte.

Deloitte webinar
Chat GPT for business problems

Evalds Urtans
evalds@asya.ai
CEO